The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 3, No. 1. Online ISSN: 2831-4794
Available at https://www.isods.org/publication/

An Overview of Facial Emotion Recognition Using Deep Learning

Vo Nhat Anh Nguyen[a], Thinh Vo Xuan[b], Tan Duy Le[a], An Mai[a], Kha Tu Huynh*[a]
[a]International University, Vietnam National University of Ho Chi Minh City
Ho Chi Minh City, Vietnam; [b] Saigon Technology University, Vietnam.

ABSTRACT

The paper presents an overview of Facial Expression Recognition methods using deep learning models that have been reviewed from published papers in reputable journals and conferences. Based on these studies, the authors have categorized the methods into five groups, including Convolutional Neural Network, Generative Adversarial Networks, Recurrent Neural Network, Transfer Learning and Other Methods and Signal-based Methods. In each group, the authors summarize prominent methods and comparative results. From these groups of methods, the authors have suggested a general process for the Facial Expression Recognition problem consisting of six steps of Problem Statement, Image preparation, Data Pre-processing, Feature extraction, Deep learning approaches and Facial expression classification. The paper provides researchers in the field of Facial Expression Recognition or newcomers to this field with an overview as well as a basis for developing in-depth research or developing solutions and models applied to related areas.

Keywords: face emotion recognition, facial expression, deep learning, CNN, GAN, RNN.

1. INTRODUCTION

Facial expressions are important in communication because they allow for the expression and inference of information about an individual's emotional state. Facial expressions are important in social interactions and are commonly used in the behavioral study of emotions. As a result, the ability to precisely discern a person's emotions is a potential area of research in many fields.

Facial Expression Recognition (FER) can aid in assessing the suitability of candidates for client-facing roles by gauging confidence levels and sincerity through emotion classification. In video game beta testing, real-time facial expression analysis provides valuable insights into players' emotional responses, aiding in game refinement. For marketing research, observing users' facial expressions during product trials enhances the understanding of consumer experiences. The application extends to AI robots, where integrating emotion recognition improves human-robot interactions, fostering more efficient and meaningful conversations. In addition, FER can also be helpful in mental health care as emotion represents a person’s mental state. Lastly, facial emotion recognition in crowd testing helps evaluate audience engagement during diverse events, offering valuable data for optimizing presentations and performances.

FER has continuously been an easy task for individuals, but completing the same task using a computer-based approach or algorithm is extremely tough and demanding. One way to increase efficiency was to employ machine learning for picture categorization. The rise of Deep Learning has yielded several promising outcomes for image classification and computer vision problems. Deep Learning outperforms regular Machine Learning approaches in terms of accuracy. For FER, multiple extensions of the Convolutional Neural Network are utilized to further improve the accuracy of emotion classification. Research on FER (Facial Expression Recognition) using deep learning models has been increasing and has become one of the topics attracting researchers. According to statistics from reputable journals and conferences over the past five years, the number of publications related to FER using deep learning models is presented in Figure 1.

*hktu@hcmiu.edu.vn

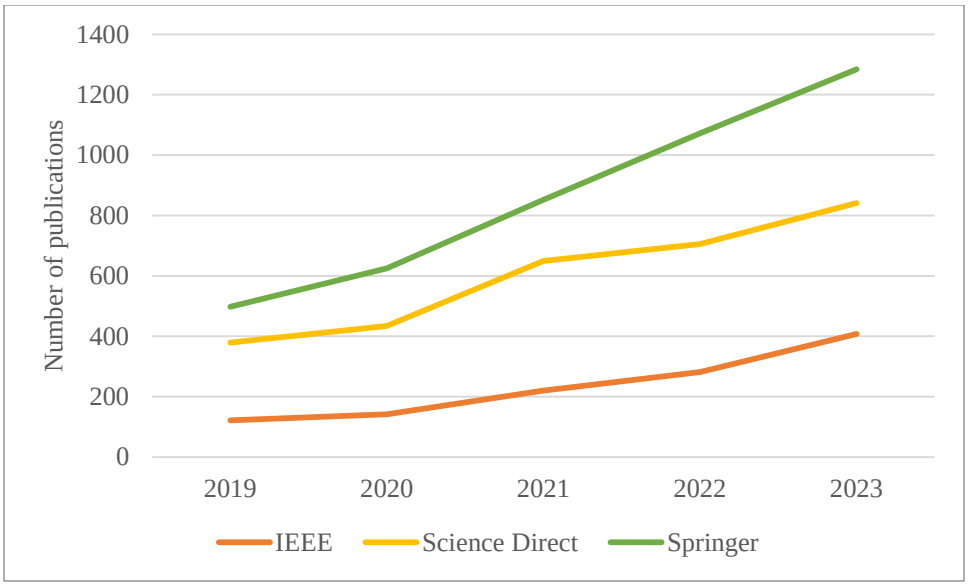


Figure 1. Statistics of scientific publications on 'FER using deep learning models' from 2019 to 2023 from IEEE Explore, ScienceDirect, Springer.

In this paper, following the introduction (section 1), the authors have selected prominent papers, categorized the methods (section 2), proposed a general process for the FER problem using deep learning models (section 3), and provided conclusions on the contributions made by the paper (section 4).

2. FACIAL EMOTION RECOGNITION USING DEEP LEARNING

In recent years, CNN has brought many effective solutions to the scientific community, among which FER methods built from CNN models have been a prominent topic. A Convolutional Neural Network's (CNN) overall design consists of numerous interwoven layers that may be adjusted to boost accuracy rates and execute specific tasks. CNN designs all contain convolutional layers, which are the essential building blocks of a CNN that help extract visual information from picture input. Some of the common Image Classification architectures are VGG-16, ResNet, DenseNet, SqueezeNet, AlexNet, Inception (GoogLeNet), and MobileNet are common Image Classification architectures. Furthermore, Haar Cascade, MTCNN (Multi-task Cascaded Convolutional Networks), and FaceNet are standard methods in FER to solve face detection issues.

FER methods using deep learning can be classified as shown in Figure 2.

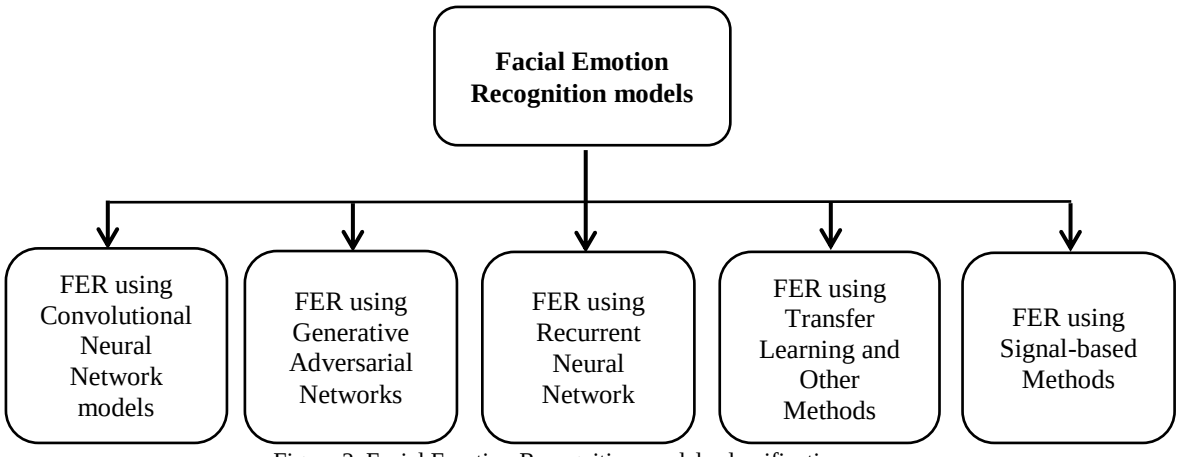


Figure 2. Facial Emotion Recognition models classification

2.1 FER using Convolutional Neural Network

In Facial Expression Recognition (FER), using Convolutional Neural Networks (CNNs) is a game changer, providing a sophisticated way to decode the complex language of human emotions hidden in facial expressions. This section will go through several CNN models in FER, demonstrating its potential to automatically extract relevant features and hierarchically learn patterns, increasing the accuracy and efficacy of emotion recognition systems.

A CNN neural network consists of convolutional layers to employ the convolution operation. Shekhar Singh and Fatma Nasoz [1] (2020) have established a CNN architecture consisting of six Convolutional Layers with ReLU activation function, three Max-Pooling, two 0.2 value Dropout layers, a flattened layer, and two Dense Layers, one with ReLU and another with SoftMax. The proposed model achieved 61.7% test accuracy without pre-processing or feature extraction techniques. Ruhi Jaiswal [2] (2020) proposes a fully convolutional neural network with residual depth-wise separable convolutions, followed by a batch normalization procedure and a ReLU activation function. In the last layer, a soft-max activation function and global average pooling are used to get a prediction. Also, HaarCascade is applied to recognize the input face, which helped the model achieve 66% accuracy. This shows that data pre-processing can improve the model performance.

However, for the same well-prepared dataset, the architecture of CNN will determine the performance of the model; a study by Jie Shao and YongShengQian [3] (2019) has proposed three kinds of CNN architecture: The first is a shallow network called the Light-CNN with a fully convolutional neural network made up of six depth wise separable residual convolution modules. The second is a dual-branch CNN that extracts classic LBP and deep learning features simultaneously. The third is a pre-trained CNN developed using the transfer learning approach to overcome a lack of training samples. The three models (Light-CNN, Dual-Branch CNN, Pre-trained CNN) were evaluated on three publicly available datasets (CK+, BU-3DFE, FER2013) and achieved an accuracy of 92.86%, 85.71%, 95.29%, 86.20%, 48.17%, 86.50%, 68%, 54.64%, 71.14% respectively. The results demonstrated that the deeper model performs better on FER, and Light-CNN could also be a good model for the FER task.

A study by Deepak Kumar Jain et al. [4] (2019) proposed a FER model with extensive pre-processing that involves image normalization, cropping to eliminate non-expression features, and intensity normalization using contrastive equalization. Feature detection focuses on key facial areas like forehead, eyebrows, eyes, cheeks, and mouth. The model architecture consists of a face detection CNN, two residual blocks, and regression DNN for valence label approximation. The results show that the combination of Fully Convolutional Networks and residual block cloud delivers a better performance of 95.23% and 93.24% for JAFFE and CK+ datasets, respectively. Besides typical CNN architecture, pre-processing, and feature extraction process, Karnati Mohan et al. [5] (2020) developed an edge descriptor based on the Law of Universal Gravitation. They propose a two-part model: an edge descriptor based on the Law of Universal Gravitation by treating each pixel in a grayscale image as a mass and a Dual Convolutional Neural Network (DCNN) for Facial Expression Recognition (FER). The model is experimented on five datasets (FER2013, JAFFE, CK+, KDEF, and RAF) with a high accuracy rate of up to 98% for JAFFE and CK+.

With the progress of technology, increasingly advanced techniques have been developed that combine several modalities and use creative fusion procedures. An example of such a method is the Min Wu et al [6] (2022) investigation. This approach integrates facial expressions and speech patterns to enhance the accuracy of identifying and interpreting changing emotions. The method utilizes the Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) technique to extract low-level dynamic characteristics from facial expressions. Additionally, it incorporates spectrograms to collect both time and frequency information in voice sounds by using Two deep convolutional neural networks (DCNNs) are used to obtain high-level semantic characteristics. The main innovation is the Two-Stage Fuzzy Fusion Strategy (TSFFS), which combines Canonical Correlation Analysis (CCA) for merging features and a Fuzzy Broad Learning System (FBLS) for merging decisions. This successfully deals with imbalances between different data types and emotional states uncertainties. The model undergoes testing on three databases: SAVEE, eNTERFACE'05, and AFEW, attaining recognition accuracies of 99.79%, 90.82%, and 50.28%, respectively. Nevertheless, this approach demonstrates constraints in intricate real-world settings, highlighting the need for more improvement.

In order to address the difficulties presented by real-world settings such as occlusion, significant head postures, and lighting fluctuations, Jae Young Choi et al. [7] (2023) use the ensemble technique. This method involves generating a collection of Deep Convolutional Neural Networks (DCNNs) and optimizing the ensemble weights using a simulated annealing (SA) process. The optimization process aims to reduce the generalized classification error by efficiently merging the outputs of several DCNNs. Different face representations such as local binary patterns and Gabor features,

together with a bagging approach, are implemented to enhance the variety among the members of the DCNN ensemble. This diversity is essential for enhancing overall performance. When tested on FER datasets such as FER2013, SFEW2.0, and RAF-DB, this technique obtained facial expression recognition accuracies of 76.69%, 58.68%, and 87.13%. However, this technique has its limits. It exhibits reduced accuracy in recognizing expressions like "disgust" and "fear," it is also constrained by computational restrictions when used in real-time applications. In response to the same issue, Jianghai Lan et al. [8] (2023) propose a novel approach called Multi-Regional Coordinate Attention Residuals (MrCAR). The process begins with using Multi-task Cascaded Convolutional Networks (MTCNN) for face identification and alignment. This involves identifying crucial areas and cropping global and local inputs, explicitly focusing on the eye and mouth regions to capture intricate details. The central component of the MrCAR model is the coordinate attention residual network (CA-Net). This network improves upon traditional residual networks by incorporating coordinate attention mechanisms. It achieves this by breaking down 2D global pooling into 1D operations, allowing for more effective directional and positional information capture. Additionally, the model utilizes multi-scale convolution to capture features at various scales. MrCAR employs Arcface Loss, rather than the conventional Softmax Loss, for classification. This choice enhances the discriminative capability of the model by increasing the distance between different classes and decreasing the distance within the same class. MrCAR had exceptional performance across many datasets, achieving an accuracy of 98.78% on CK+, 99.09% on JAFFE, 74.50% on FER2013, and 88.26% on RAF-DB. However, it continues to encounter constraints when dealing with the diverse range of natural picture variations, incorrect labeling, areas that are not faces, and instances of occlusion. Although the model has successfully decreased the number of incorrect classifications for negative emotions, it still needs help distinguishing tiny differences between these expressions. This suggests that more improvements are necessary. Ensemble approaches and two-stage algorithms are particularly successful in facial expression recognition (FER). The assertion is supported by the research conducted by Sun-Hee Kim et al. (2022) [9], who proposed an innovative two-stage methodology for emotion identification that is crucial for human-machine interaction systems. The Tiny Face Detector recognizes and extracts face areas from each frame in a video sequence. The convolutional neural network (CNN) extracts essential properties from these facial areas, collecting low-level, mid-level, and high-level characteristics. In the second step, the extracted features are treated as temporal data to analyze the changes in facial expressions across successive frames. This uses a Multi-Level Convolutional Neural Network (MLCNN) that integrates connections from different levels to use local and global characteristics. To improve this method, the authors suggest employing an ensemble model that combines three different versions of MLCNNs. Each variation utilizes distinct kinds of mid-level connections to capture more complex data and boost the overall accuracy. The MLCNN model, which was trained using the FER2013 dataset, obtained an accuracy of 73.03%. In comparison, the ensemble model achieved a higher accuracy of 74.09%, surpassing the performance of other advanced approaches. The ensemble model outperformed the single architecture of the EmotiW 2017 competition winner on the AFEW 7.0 dataset, demonstrating its efficacy in dealing with video-based emotion identification. While this method effectively addresses several facial expression recognition (FER) problems in natural settings, the intricate nature of the ensemble model requires substantial processing power, which limits its use in real-time applications. Additionally, the issue of imbalanced data, where some emotions have less samples, remains a significant worry.

Another notable strategy involves hierarchical feature extraction. In their study, Ying He et al. [10] (2023) propose a multi-layer feature recognition algorithm, HFT-CNN, based on a three-channel convolutional neural network. This algorithm aims to overcome traditional CNNs' limitations in facial expression recognition, specifically in terms of inadequate feature extraction and low recognition rates. The HFT-CNN architecture has three distinct channels that autonomously train on distinct facial regions: the whole face, the eyes and eyebrows, and the lips. The use of higher-order tensor singular value decomposition (HOSVD) in each channel optimizes convolution kernels, eliminating duplication and improving the efficiency of feature extraction. Additionally, the model utilizes mean and variance management of convolutional kernels to distinguish between low-frequency (contour) and high-frequency (detail) information. The characteristics extracted from the three channels are combined to give the ultimate expression categorization. This fusion procedure aims to increase the importance of more significant areas of the face, such as the lips, compared to the eyes. The CK+ and JAFFE datasets experiments showed that HFT-CNN produces superior recognition rates compared to standard deep neural networks and state-of-the-art approaches, while having fewer network layers. Nevertheless, the research recognizes that the model's effectiveness relies heavily on the proper extraction of face areas and may experience a decline in performance when dealing with less controlled or diverse input data.

Furthermore, within the realm of hierarchical feature extraction, the CHC (Cross-Hierarchy Contrast) framework exists, visualized in a publication by Xuemei Wu et al. in 2023 [11]. This methodology effectively improves the performance of

facial expression recognition (FER) by incorporating contrastive learning into a deep learning architecture based on convolutional neural networks (CNNs). The CHC framework utilizes the ResNet architecture for hierarchical feature extraction, while including a contrastive learning mechanism to effectively capture intricate and essential data about face expressions. The CHC module compares hierarchical characteristics to increase similarities in standard features and reduce differences in distinct features. The fusion network combines these characteristics, resulting in a resilient representation for categorization. FER-CHC was tested on six FER datasets, namely CK+, FER2013, FER+, RAF-DB, AffectNet, and JAFFE. It demonstrated state-of-the-art performance, achieving accuracies of 99.75%, 92.9%, 74.68%, 90.81%, 90.06%, and 66.31% on these datasets, respectively.

Attention mechanisms have emerged as a vital component in increasing the performance of convolutional neural networks (CNNs) for facial expression recognition (FER). Attention mechanisms that enable the model to concentrate on important information boost recognition accuracy. Yinggang He's work [12], published in 2023, used the model to analyze grayscale face expression photos by using three branches, each specifically engineered to capture multi-scale characteristics from various network levels. The branches of the model use depth-wise separable convolutions, which split conventional convolutions into depthwise and pointwise convolutions. This division dramatically reduces the computing load and number of parameters required, while maintaining high performance. Moreover, the Convolutional Block Attention Module (CBAM) is included in each branch to enhance important characteristics and reduce the impact of less significant information using channel and spatial attention processes. The network incorporates Rectified Linear Unit (ReLU) activation functions and batch normalization layers to expedite the training process and enhance the stability of the model. The results from the three branches are combined using element-wise summation, and then a global average pooling operation is used to provide a concise feature representation. The ultimate feature vector is inputted into a fully connected layer, which employs a softmax function to forecast the probability distribution across various facial expression categories. By conducting thorough experiments on FER2013, FERPLUS, and CK+ datasets, the model obtained identification accuracies of 69.49%, 84.633%, and 99.39% respectively. These results indicate substantial enhancements compared to conventional single-branch CNNs and other cutting-edge techniques. Nevertheless, the challenge of differentiating tasks based on facial expressions with comparable characteristics remains owing to a need for adequate training samples for distinct emotions.

In addition to 2D convolutional techniques, Zhifei Han et al. [13] (2024) use the AM-3D CNN model. This sophisticated model combines a three-dimensional Convolutional Neural Network (3D CNN) with an attention mechanism to improve emotion perception. The 3D CNN method captures emotional expressions' spatial and temporal aspects from multimodal data, such as facial expressions and body movements. This integration enhances the model's ability to manage the intricate and multifaceted character of emotions successfully. The attention mechanism enhances the performance of the 3D CNN by enabling the model to choose to concentrate on significant characteristics within the emotional data, enhancing accuracy and resilience. Compared to other models like classic 2D CNN, SVM, RNN, and LSTM, the AM-3D CNN model outperforms them in distinguishing different emotions. It achieves accuracy rates of 97% for anger, 91% for disgust, and 94% for happiness. However, the model's performance differs depending on the specific emotion categories, and it may have difficulties when dealing with emotions that could be more well expressed. Moreover, the need for top-notch multimodal data and substantial computing resources for training and deployment might impose limitations. Future research should prioritize increasing the dataset, integrating more modalities including text and audio, and improving the attention mechanism to improve the model's ability to generalize and be practical.

These CNN-based techniques demonstrate the ongoing progress and ingenuity in recognizing face expressions. They tackle distinct obstacles and enhance performance on varied datasets. Every technique has its own advantages and disadvantages, which help develop more reliable and precise FER systems continuously, detailed as in Table 1.

Table 1 Summary of CNN-Based Facial Expression Recognition (FER) Approaches

Paper	**Dataset**	**Accuracy**	**Key Technique**	**Limitation**
Facial Expression Recognition with Convolutional Neural Networks [1]	FER2013	61.70%	CNN with ReLU activation, Haar Cascade classifier for face detection, histogram equalization	Overfitting, low test accuracies for certain emotions

Facial Expression Classification Using Convolutional Neural Networking and Its Applications [2]	FER2013	61.70%	Normalization, data augmentation	Computational complexity, performance in varying conditions
Three Convolutional Neural Network Models for Facial Expression Recognition in the Wild [3]	CK+, BU-3DFE, FER2013	95.29%, 86.50%, 71.14%	Light-CNN, Dual-Branch CNN, Pretrained CNN (ResNet101)	Inefficiency of traditional features, computational intensity
Extended deep neural network for facial emotion recognition [4]	CK+, JAFFE	93.24%, 95.23%	DNN with deep residual blocks, image normalization	Decreased recognition rates for non-frontal faces, varying lighting conditions
Facial Expression Recognition Using Local Gravitational Force Descriptor Based Deep Convolution Neural Networks [5]	FER2013, JAFFE, CK+, KDEF, RAF	77.10%, 95.63%, 97.99%, 96.36%, 82.33%	Local gravitational force descriptors, dual-branch DCNN	Overfitting, high computational cost, lower performance on real-world datasets
Two-Stage Fuzzy Fusion Based-Convolution Neural Network (TSFFCNN) [6]	SAVEE, eNTERFACE'05, AFEW	99.79%, 90.82%, 50.28%	Integration of Canonical Correlation Analysis (CCA) and Fuzzy Broad Learning System (FBLS)	The reduction of accuracy in complex real-world environment
Combining Deep Convolutional Neural Networks with Stochastic Ensemble Weight Optimization for Facial Expression Recognition in the Wild [7]	FER2013, SFEW2.0, RAF-DB	76.69%, 58.68%, 87.13%	Ensemble of DCNNs with ensemble weights optimized using simulated annealing	Lower accuracy for disgust and fear, computational efficiency constraints
Expression Recognition Based on Multi-Regional Coordinate Attention	CK+, JAFFE, FER2013, RAF-DB	98.78%, 99.09%, 74.50%,	Multi-Regional Coordinate Attention Residuals (MrCAR)	Mislabeling non-face regions, struggles with

Residuals [8]		88.26%		subtle variations in negative expressions
Facial Expression Recognition Using a Temporal Ensemble of Multi-Level Convolutional Neural Networks [9]	FER2013, AFEW 7.0	74.09% (FER2013)	Two-stage technique with MLCNN and Tiny Face Detector	Significant computational resources required, unbalanced data challenge
FER-CHC: Facial expression recognition with cross-hierarchy contrast [11]	CK+, JAFFE, FER+, RAF-DB, AffectNet, FER2013	99.75%, 92.9%, 74.68%, 90.81%, 90.06%, 66.31%	Cross-hierarchy contrast with ResNet and contrastive learning	Potential noise from in-the-wild data, increased training complexity
Emotion Recognition and Management in the Tourism Industry During Emergency Events Using Improved Convolutional Neural Network [13]	Bimodal Face and Body Gesture Database (FABO)	88.81%	3D CNN with attention mechanism	Relies on high-quality multimodal data, significant computational resources

2.2 FER using Generative Adversarial Networks

GAN is an unsupervised learning model combining Generative and Discriminant networks. It has been effectively utilized in image synthesis to produce realistic pictures such as images and videos. The GAN-based model is helpful for data augmentation and favorable to expression recognition with invariant posture and identity. There exist numerous types of GAN models utilized in the domain of facial expression recognition, each designed to address different aspects of the task.

Ziyang Han et al. [14] (2019) suggested a face-merged GAN that uses two generators and one discriminator. With this approach, the confrontation power is enhanced, and so is the model's performance. The initial generator utilizes a pre-trained Resnet50 encoder to extract key facial features and produce a synthetic face. The encoded information is then divided into upper and lower parts, processed by separate partial decoders, and merged to generate the first frontal face. The second generator refines this output by incorporating multi-scale features and high-level information, enhancing photorealism. Based on the competition with the discriminator (employs Wasserstein distance), the second generator synthesizes the final frontal face. To reduce model complexity, the face merging GAN encoded the original image once using the pre-trained network and then shared the recovered features across two decoders. Therefore, only the parameters of two decoders and one discriminator must be changed during training. Their experiments delivered a performance of 91.8% accuracy and showed that the frontal face synthesized by the face-merging GAN kept the face's identity and realism better than other GANs.

Building on the idea of enhancing GAN performance, Dharanya V et al. [15] (2021) have proposed a model based on the Auxiliary Classifier GAN that could create ten expressions from input images. To alleviate the subject dependence issue, they trained the model individually and created all the phrases above for a single person, allowing the discriminator to

detect expressions. The U-NET architecture implements the generator, whereas the recognizer extracts features using a capsule network. The suggested model was assessed on the ADFES-BIV dataset and delivered an overall classification accuracy of 93.4%. The proposed model was evaluated on the ADFES-BIV dataset with an overall classification accuracy of 93.4%. Instead of generating images, a novel feature separation model by Lie Yang et al. [16] (2020) can distinguish expression-related data from unrelated ones. The model involves a two-stage process: training the exchange-GAN feature separation and then using the separated features for the FER task. Exchange-GAN is trained by partial feature exchange, and various constraints effectively separate expression-related and expression-independent features with high purity. To mitigate the negative impact of identity-related characteristics, we disregard expression-independent data and focus on expression-related features while conducting FER. These two parts of the process can help the model effectively overcome the interference of identity-related features. The model was then evaluated on three experimental databases, Oulu-CASIA, CMU Multi-PIE, and FACES, and achieved competitive accuracy of 91.08%, 95.24%, and 86.33%, respectively. Ali et al. [17] (2021) also used the feature disentanglement approach and identity-invariant expression representation learning, a similar principle. The researchers introduced a unique technique called Transfer-based Expression Recognition Generative Adversarial Network. The Transfer-based Expression Recognition Generative Adversarial Network acquired the ability to separate facial features by extracting distinctive characteristics from one image and transferring them to identify the second image. To improve the process of disentangling features and acquire a representation of facial expressions invariant to identity, the researchers used expression consistency loss and identity consistency loss. The model was tested on four public facial expression datasets (CK+, Oulu-CASIA, MMI, and BU-3DFE), resulting in accuracies of 98.47%, 89.14%, 84.83%, and 74.69%, respectively. Due to the ambiguous nature of distinguishing between expressions, even little modifications in a specific region of the human face may result in changes to facial expressions. This is because the same facial expression can exhibit different features from different angles. To address the common issues of overlapping and blurring that occur when generating images, especially in areas with intricate facial characteristics such as the eyes and lips, Yifan Xia et al. [18] (2022) have introduced the Local and Global Perception Generative Adversarial Network (LGP-GAN). This method aims to enhance the quality of facial expression synthesis by specifically targeting both local and global facial features. The LGP-GAN utilizes a two-stage cascaded architecture. In the first stage, local networks are employed to generate detailed local regions of the face, specifically the eyes and mouth. In the second stage, a global network is used to generate the overall facial structure based on the outputs from the local networks. This approach guarantees that the generated facial expressions retain high accuracy and sharpness, especially in areas crucial for distinguishing distinct emotions. Experimental findings, assessed on the Radboud Faces Database (RaFD), indicate that LGP-GAN surpasses state-of-the-art techniques like StarGAN, AttGAN, and GANimation, obtaining higher performance in terms of Inception Score (IS) and Fréchet Inception Distance (FID).

In the realm of improved GAN, Zhe Sun et al. [19] (2022) provide a new strategy in the field of enhanced GANs, which combines an improved conditional generative adversarial network (im-cGAN) with a discriminative deep fusion (DDF) technique. The im-cGAN is used to produce labeled expression samples by training on face pictures with action units (AUs), therefore augmenting the training dataset. The DDF technique combines global features acquired via a global-based module with local features acquired through a region-based module to achieve a complete feature representation. A discriminative loss function (D-loss) is specifically formulated to optimize the separation between different classes and reduce the distances within each class, resulting in enhanced discrimination of the combined data. This technique has been shown successfully by extensive tests on four public datasets: JAFFE, CK+, Oulu-CASIA, and KDEF. The approach attained an accuracy of 98.37% on the JAFFE dataset, 98.10% on the CK+ dataset, 93.34% on the Oulu-CASIA dataset, and 98.30% on the KDEF dataset. The suggested technique demonstrates a higher level of performance in comparison to many cutting-edge methods, suggesting its potential to enhance the accuracy of facial emotion identification. Fangzheng Huang et al. [20] (2024) propose a solution to enhance recognition performance in various imaging conditions. Their approach involves using a novel model called Cyclic-Style GAN (CS-GAN), which combines the features of Style-GAN 3 and Cycle-GAN. This model aims to overcome the differences between NIR and VIS images. The CS-GAN model is specifically developed to enhance the accuracy of face recognition systems by generating realistic visible (VIS) pictures from near-infrared (NIR) data. The architecture comprises two distinct Generative Adversarial Networks (GANs): one dedicated to generating visible (VIS) pictures (GV) and another focused on generating near-infrared (NIR) images (GN). The mapping network receives latent vectors (z) as input and transforms them into a latent subspace (W) using two fully connected layers. This process separates the features to enhance the ability to manipulate the resulting pictures. The synthesis network enhances the mapped latent vectors by iterating over 14 synthesis layers, gradually increasing the picture resolution and including finer features. The VIS synthesis network incorporates an extra to-RGB layer to provide color pictures, while the NIR synthesis network produces grayscale

images. The discriminator in each GAN assesses the genuineness of the produced pictures by transforming them into feature matrices and using a logistic loss function. An essential element of the approach is cyclic subspace learning, which incorporates a hidden loss in the latent subspace (W) to ensure feature consistency, style loss to maintain visual quality, and cycle consistency loss to keep identity and detail throughout translations. The CS-GAN undergoes pre-training on the FFHQ-U dataset, an unaligned version of the FFHQ dataset. This pre-training phase helps the CS-GAN to learn face features initially. After that, the CS-GAN is fine-tuned on the CASIA NIR-VIS 2.0 database to specialize in translating between near-infrared (NIR) and visible (VIS) images. This fine-tuning process allows the CS-GAN to adapt to the unique characteristics and variations of the NIR-VIS dataset. The CS-GAN achieves a cutting-edge Rank-1 accuracy of 99.6% on the CASIA NIR-VIS 2.0 database. Both researches have successfully developed and introduced innovative GAN models. However, a common challenge they face is the computational complexity of their techniques, which may restrict their practical use in real-time situations.

Facial expression recognition (FER) typically prioritizes the identification of the six fundamental emotions outlined by Ekman, namely anger, contempt, fear, happiness, sorrow, and surprise, in addition to a neutral category. Recent developments in the area have broadened the range of applications for GANs, including specific tasks like identifying sexual facial expressions. An exemplary contribution to this field is the investigation conducted by Abhishek Gangwar and colleagues [21] (2023). Triple-BigGAN tackles the task of Sexual Facial Expression Recognition (SFER), which is different from typical FER systems that classify expressions into the well-recognized seven emotions. This framework presents a novel dataset called SEA-Faces-30k, which consists of 30,817 photos categorized into three unique groups: erotic, suggestive-erotic, and non-erotic. Triple-BigGAN employs a semi-supervised learning approach that integrates a class-conditional generator, discriminator, and classifier. The generator creates synthetic pictures of excellent quality using class labels, while the discriminator distinguishes between actual and synthetic image-label combinations. Additionally, the classifier acquires the ability to classify expressions using labeled and unlabeled data effectively. This approach not only attains a high level of accuracy in classifying (93.59% on the SFER test), but it also produces authentic facial emotions that augment the variety of the training data.

Generative Adversarial Networks (GANs) have been extensively used in facial expression recognition (FER) to improve the precision and resilience of emotion identification models. GANs excel at producing authentic facial expressions, making them valuable for enhancing training datasets and enhancing model efficacy. Nevertheless, subsequent studies have emphasized the susceptibility of FER models to adversarial assaults, whereby introducing undetectable changes to input pictures may lead to the misclassification of emotions. Daeha Kim and colleagues [22] (2024) have introduced the Facial Attention-aware Adversarial Training (FAAT) technique in response to this issue. The FAAT method enhances the resilience of FER models by prioritizing important face areas using class activation maps (CAM). It also incorporates frequency-based regularization to match the frequency characteristics of clean and hostile pictures. This method guarantees that FER models can preserve their excellent performance even when faced with adversarial noise. By incorporating approaches such as Fast Adversarial Attack Transfer (FAAT) into Generative Adversarial Network (GAN)-based Facial Expression Recognition (FER) models, the reliability and security of applications may be enhanced. This is especially important in sensitive real-world situations where the potential for adversarial assaults is a substantial concern. The experimental findings indicate that FAAT dramatically improves the robustness of FER models against several types of assaults, such as Projected Gradient Descent (PGD) and Fast Gradient Sign Method (FGSM), resulting in performance improvements of up to 18% in some scenarios.

GANs provide significant advantages in generating high-quality target samples, including assisting in identifying face expressions invariant to posture and expanding the diversity of training datasets. The limited availability of public datasets, including facial expressions, emphasizes the need to use pictures produced by GANs to improve the performance of models. Models such as the face-merged GAN, Auxiliary Classifier GAN, im-cGAN, and Cyclic-Style GAN, together with specific applications like Triple-BigGAN and FAAT, tackle different issues related to Facial Expression Recognition (FER). These developments not only boost the accuracy of detecting emotions by eliminating interference from irrelevant input but also improve the strength and dependability of emotion recognition systems, hence facilitating the development of more secure and practical applications in real-world scenarios. The summary of GAN-Based Facial Expression Recognition (FER) Approaches with their performance as presented in Table 2.

Table 2 Summary of GAN-Based Facial Expression Recognition (FER) Approaches

Paper	**Dataset**	**Accuracy**	**Key Technique**	**Limitation**
Face Merged Generative Adversarial Network with Tripartite Adversaries [14]	KDEF, CMU Multi-PIE	91.8% (KDEF)	FM-GAN with two generators and one discriminator	Reliance on pre-trained networks, high computational resources needed
Facial Expression Recognition through person-wise regeneration of expressions using AC-GAN [15]	ADFES-BIV	93.4%	AC-GAN with U-Net and Capsule Networks	Sensitivity to lower intensity expressions, reliance on controlled datasets
A Novel Feature Separation Model Exchange-GAN for Facial Expression Recognition [16]	Multi-PIE, FACES, Oulu-CASIA	91.08%, 95.24%, 86.33%	Exchange-GAN for separating expression-related features from identity-related features	Requires extensive experimentation for feature segmentation ratio, performance degrades with low-definition images
Facial Expression Recognition by Using a Disentangled Identity-Invariant Expression Representation (TER-GAN) [17]	CK+, Oulu-CASIA, MMI, BU-3DFE	98.47%, 89.14%, 84.83%, 74.69%	TER-GAN with adversarial and identity consistency loss	Reliance on synthetic image quality, need for extensive training data, potential overfitting in small datasets
Local and Global Perception Generative Adversarial Network for Facial Expression Synthesis [18]	RAFD-database	1.31 (IS), 11.88 (FID)	Two-stage cascaded GAN with local and global network	Limited performance in natural and uncontrolled conditions
A discriminatively deep fusion approach with improved conditional GAN (im-cGAN) for facial expression recognition [19]	JAFFE, CK+, Oulu-CASIA, KDEF	98.37%, 98.10%, 93.34%, 98.30%	im-cGAN with Discriminative Deep Fusion (DDF)	Computational complexity limits applicability in real-time video-based tasks
Cyclic Style Generative Adversarial Network for	CASIA NIR-VIS 2.0	99.6%	CS-GAN with Style-GAN and Cycle-GAN integration	Challenges with resolution and edge-

Near Infrared and Visible Light Face Recognition [20]				blurring, needs enhancement for real-time applications
Triple-BigGAN: Semi-supervised generative adversarial networks for image synthesis and classification on sexual facial expression recognition [21]	SEA-Faces-30k	93.59%	Triple-BigGAN with class-conditional generator, discriminator, and classifier	Relies on new specialized dataset, requires further validation across diverse attributes
Towards the adversarial robustness of facial expression recognition: Facial attention-aware adversarial training [22]	-	Up to 18% improvement	FAAT with class activation maps and frequency-based loss regularization	Potential performance degradation under severe illumination and varying head poses

2.3 FER using Recurrent Neural Network

RNNs, a deep learning approach for describing dynamic temporal behavior, distribute weight throughout their network, enabling the output of a neuron to be fed back into the same neuron. This series network topology is particularly effective with time series data, maintaining data dependence and allowing for sophisticated temporal modeling.

Various studies have explored the application of RNNs in facial expression recognition (FER). For instance, Rokkones et al. [23] (2019) approach the problem of human-machine interaction with a novel method of combining edge features in facial expressions with a recurrent neural network (RNN) to predict different emotions. The feature extraction approach includes Kirsch edge detection and the Local Directional Strength Pattern (LDSP). Kirsch edge detection uses eight directional masks to highlight edges surrounding a pixel. LDSP - an enhancement on Local Directional Pattern (LDP) provides a 6-bit representation by considering strengths in eight directions, resulting in more robust edge patterns. For modeling expressions, a Recurrent Neural Network (RNN) is employed to handle sequential data, predicting the next step in a sequence based on temporal observations. The Local Directional Strength Pattern RNN had a classification accuracy of 73.60% for CK+.

In a different approach, Hanyu Liu et al. [24] (2020) proposed a model to recognize emotion in wild Videos by combining a Deep Residual Network (ResNet) and Bidirectional Recurrent Network (RNN) with a Long-Short-Term Memory Unit (BLTSM). The model architecture starts with ResNet-101 and a Convolutional Block Attention Module (CBAM) to effectively extract visual data from each frame. ResNet-101 acts as the feature extraction backbone, with CBAM added after each residual block to integrate both channel and spatial attention, boosting the model's emphasis on face expression areas. To recognize the continuous character of facial emotions bidirectional long short-term memory network (BLSTM) is employed to capture temporal relationships, which recognize the continuous character of facial emotions. The BLSTM processes time data, and fully linked layers are then used in the classification module to categorize the collected characteristics into seven facial expression classes. The model achieved 67% accuracy on the Aff-Wild2 dataset, which is 10.8% higher than the 30% baseline. While RNN can accurately recognize FER from video sequences, it is prone to gradient explosion and disappearance.

Rodriguez et al. [25] (2022) provide more evidence of the effectiveness of LSTM networks in Facial Expression Recognition (FER) via their approach for identifying pain expressions. Their research utilizes LSTM networks to capture the temporal relationships of face emotions in video sequences. The method entails preprocessing video frames by

aligning, cropping, and frontalizing faces using landmarks. The LSTM network utilizes its ability to record the temporal aspect of facial emotions, allowing the system to precisely identify pain by evaluating the evolution of facial movements. Using a leave-one-subject-out cross-validation methodology, this approach gets an impressive area under the curve (AUC) score of 93.3% on the state-of-the-art UNBC-McMaster Shoulder Pain Expression Archive Database. Furthermore, it exhibits competitive performance when tested on the Cohn Kanade+ facial expression database, suggesting its capacity to perform well on many facial expression recognition tasks. Nevertheless, several constraints such as the susceptibility to data imbalance and the dependence on accurate facial landmark recognition impact the system's effectiveness in diverse real-life situations. Manalu and Rifai [26] (2024) created a hybrid CNN-RNN model to identify human emotions from video data. They combined CNN to extract spatial features and RNN, namely LSTM, to manage temporal dependencies. Their research used the Emotional Wearable Dataset 2020, which encompasses feelings such as amusement, excitement, amazement, and like, going beyond the conventional seven emotions. Three models were created: a bespoke CNN-RNN model, MobileNetV2-RNN, and InceptionV3-RNN. The bespoke CNN-RNN model attained an accuracy of 63%, the MobileNetV2-RNN reached 59%, and the InceptionV3-RNN earned the maximum accuracy of 66%. The LSTM component of their approach enhances the temporal interdependence of video sequences. Their bespoke CNN-RNN model attained a precision rate of 63%, while the InceptionV3-RNN model acquired a precision rate of 66%. This study highlights the significance of Recurrent Neural Networks (RNNs) in improving the capacity to identify a broader spectrum of emotions in Facial Expression Recognition (FER). It showcases the promise and difficulties of using RNNs in this field, including the issues of overfitting and distinguishing closely related emotions.

The efficacy of FER systems, especially in real-time applications, may be significantly impeded by prediction delay. In order to tackle this problem, Baddar et al. [27] (2022) devised a technique that utilizes Long Short-Term Memory (LSTM) networks to make real-time predictions using incomplete sequences. This method improves the real-time response speed by including characteristics that minimize the impact of the subject's appearance on the spatio-temporal data stored by the LSTM. This technique greatly enhances the early prediction of subtle expression frames. It increases overall identification accuracy by using a fixed sequence to estimate and reduce appearance bias and creating dense labels at the frame level to regulate the training process. Their method achieved recognition rates of 82.97% on the MMI dataset, 86.25% on the Oulu-CASIA dataset, 66.80% on the CASME II dataset, and 51.17% on the AFEW dataset. Although the technique has shown promising outcomes, it is essential to note its limitations. These include the high computing requirements for estimating frame-level intensity and its vulnerability to changes in lighting and backdrop. Additionally, the method assumes that the starting frame represents a neutral expression.

In addition to the other approaches, Tong Zhang et al. [28] (2019) developed the Spatial-Temporal Recurrent Neural Network (STRNN) to combine spatial and temporal information to recognize emotions from EEG data and facial expressions. This framework utilizes a multidirectional Recurrent Neural Network (RNN) layer to capture long-range spatial dependencies and a bi-directional temporal RNN layer to capture temporal dependencies. Additionally, it incorporates sparse projection to increase its discriminative capabilities. The STRNN demonstrated remarkable performance, achieving an accuracy of 89.50% on the SEED EEG dataset and 95.4% on the CK+ dataset, surpassing other cutting-edge techniques. Nevertheless, the intricacy and need for abundant training data impose specific constraints, necessitating more refinement to address real-world discrepancies better.

Although RNNs show promise in reliably identifying facial emotions from video sequences, they face obstacles such as gradient explosion and disappearance. Each approach demonstrates the growing methodologies and breakthroughs in using RNNs for facial expression detection, highlighting their ability to process dynamic and sequential data properly. The summary of RNN-Based Facial Expression Recognition (FER) Approaches is presented in Table 3.

Table 3 Summary of RNN-Based Facial Expression Recognition (FER) Approaches

Paper	**Dataset**	**Accuracy**	**Key Technique**	**Limitation**
Facial Expression Recognition Using Robust Local Directional Strength Pattern Features	CK+	73.6%	Local Directional Strength Pattern (LDSP) features combined with Recurrent Neural	Difficulty in distinguishing between similar expressions like anger and sadness, misclassification of less

and RNN [23]			Network (RNN)	represented classes like contempt
Facial Expression Recognition for In-the-wild Videos [24]	Aff-Wild2, AffectNet, RAF-DB	66.9%	101-layer Deep Residual Network (ResNet-101) with Convolutional Block Attention Module (CBAM) and BLSTM	Reliance on accurate face detection and alignment, potential overfitting, sensitivity to variations in lighting and occlusion
Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification [25]	UNBC-McMaster, CK+	93.3% AUC (UNBC-McMaster)	LSTM for temporal dependencies in video sequences	Sensitivity to data imbalance, reliance on high-quality facial landmark detection
Detection of Human Emotions through Facial Expressions Using Hybrid CNN-RNN Algorithm [26]	Emotional Wearable Dataset 2020	63% (custom), 59% (MobileNetV2-RNN), 66% (InceptionV3-RNN)	Hybrid CNN-RNN (LSTM) for spatial and temporal features	Overfitting, difficulty in differentiating closely related emotions
On-the-Fly Facial Expression Prediction Using LSTM Encoded Appearance-Suppressed Dynamics [27]	MMI, Oulu-CASIA, CASME II, AFEW	82.97%, 86.25%, 66.80%, 51.17%	LSTM with appearance-suppressed dynamics	Computational intensity, sensitivity to lighting and background variations, initial neutral frame assumption
Spatial-Temporal Recurrent Neural Network for Emotion Recognition [28]	SEED, CK+	89.50%, 95.4%	Multi-directional RNN, Bi-directional temporal RNN	Complexity, extensive training data, performance gain varies with dataset

2.4 FER using Transfer Learning and Other Methods

Although deep learning models have consistently surpassed traditional machine learning methods in performance, they still need significant improvements due to the extensive demand for annotated training images, resulting in expensive and time-consuming operations. Transfer learning addresses this limitation. This approach involves reusing pre-trained models initially trained on extensive image classification tasks to address new and related problems with less data and reduced complexity.

Chowdary et al. [29] (2021) used transfer learning to train pre-existing networks, including ResNet50, VGG19, Inception V3, and MobileNet, for emotion recognition. These models served as basis architectures for feature extraction, with the pre-trained CNNs' fully connected layers removed and new fully connected layers added to meet the unique needs of the emotion classification job. The training technique used the CK+ dataset to demonstrate the potential enhancement of emotion identification across seven distinct emotional states via transfer learning. Consequently, they

attained a remarkable mean accuracy of 96%. Shamoil Shaees et al. [30] (2020) introduced a hybrid classifier that combines transfer learning with a classification method called Support Vector Machine (SVM). The model test was conducted on three public datasets, including NVIE and CK+ expressions, achieving accuracy rates of 99.3% and 98.3%, respectively. Cross-modal Dynamic Transfer learning (CDaT), proposed by Soyeon Hong et al. [31] in 2024, is a technique that tries to enhance model performance using pre-existing information, similar to transfer learning. CDaT stands out by prioritizing the dynamic integration and adaptation of several modalities, including text, audio, and visual input, for multimodal emotion identification. CDaT employs uni-modal masking and cross-modal representation transfer learning to remove low-confidence modalities and augment them with high-confidence modalities selectively. The process comprises two crucial phases: mismatched modality filtering and dynamic bidirectional information transmission. The researchers evaluated the performance of CDaT on four advanced fusion models (Naive et al., MISA, and TAILOR) using the CMU-MOSEI and IEMOCAP datasets. The findings demonstrated notable enhancements, as CDaT exhibited a 4.74% rise in F1 score compared to the TFN model and an 11.03% increase compared to the TAILOR model on the IEMOCAP dataset. Nevertheless, a drawback of this strategy is its reliance on dependable confidence measures for cross-modal transfer learning, which may not consistently provide positive results when applied to diverse datasets, as seen by the differing performance improvements observed between CMU-MOSEI and IEMOCAP. In addition, the Naive Fusion model exhibited inadequate learning performance since it did not have a design that addressed the disparity between different modalities.

Convolutional Neural Networks (CNNs) have made great strides in facial expression detection. However, various ways exist to integrate face graphs with other technologies to tackle issues and enhance recognition accuracy. Byoung Chul Ko et al. [32] (2023) introduce a new method for facial expression recognition (FER) under challenging conditions by integrating face graphs and an attention mechanism. This approach generates a graph of facial features by choosing action units (AUs) that are important for facial expressions. It then uses a graph convolutional network (GCN) to identify these expressions accurately. The face picture is first fed into ResNet18 to extract feature maps. These feature maps are then processed using a trilinear attention module, which generates channel-wise attention maps. These attention maps emphasize the areas of the face that are related to facial muscle movements. Facial regions, including the eyes, mouth, and chin, are isolated from the rest of the face using predetermined facial markers. The purpose of these patches is to provide node characteristics, which are then used to construct a face network. In this graph, nodes represent facial patches, and edges reflect their geometric connections. The node characteristics are inputted into a two-layer Graph Convolutional Network (GCN), which captures the relationships between nodes and generates a compact feature vector representing the face graph. The vector is then identified using a multi-layer perceptron (MLP) to ascertain the face expression. The approach has shown exceptional performance on both controlled laboratory datasets and uncontrolled field datasets. The research by Xu Xu, Zhou Ruan et al. [33] (2020) combines facial landmark identification with graph convolutional neural networks (GCNNs) by combining face graphs. The procedure starts by using the DLIB toolbox for accurate landmark identification, which utilizes a collection of regression trees. The landmarks are shown as nodes in a graph, where the edges reflect the strength of connections, representing the geometric correlations between face characteristics. The Graph Convolutional Neural Network (GCNN) analyzes the graph-structured data using a network design with nine layers. These layers consist of seven convolutional layers and two fully connected layers. Optimization involves training the network over 400 epochs using stochastic gradient descent with momentum. By conducting experiments on datasets such as JAFFE, FER2013, and CK+ and using data augmentation methods such as rotation, shifting, scaling, noise addition, contrast adjustment, and color jittering, the approach achieves a remarkable identification accuracy of up to 95.85%. FaceTopoNet combines face graphs with dynamic topology learning, expanding the existing idea. In their study published in 2023 [34], Mojtaba Kolahdouzi and colleagues developed a model called FaceTopoNet. This model generates a fully connected network using facial landmarks and then optimizes the weights of the edges using an evolutionary method. The optimized network yields a minimum-cost spanning tree, which produces sequences of face nodes processed by two distinct streams. The first stream, the structure stream, focuses on landmark placements. The second stream, called the texture stream extracts local patches surrounding landmarks. These streams are fused using an attention-based strategy, producing robust FER performance. FaceTopoNet demonstrates exceptional precision on many datasets, such as AffectNet (57.3%) and FER2013 (75.8%), surpassing all existing methods. Nevertheless, the three models have similar drawbacks, including susceptibility to occlusions and variations in position, difficulties with imbalanced data, computational complexity, and limited capacity to generalize across diverse datasets.

Another significant advancement in FER methodologies is the application of transformer models. The work of Ning Sun et al. [35] (2023) showcases an outstanding method. The authors provide a new deep neural network model dubbed the Appearance and Geometry Transformer (AGT) that utilizes self-attention for Facial Expression Recognition (FER) in

complex real-world scenarios. The AGT combines two pathways: the Image Transformer Pathway (ITP), which employs the DeiT model to extract visual characteristics from facial image patches, and the Graph Transformer Pathway (GTP), which utilizes multi-scale graph-based self-attention to capture spatial characteristics from facial landmarks. The model integrates a distillation process to enhance training effectiveness with a limited number of examples. The usefulness of the AGT has been shown by evaluations on the RAF-DB and FERplus databases, obtaining accuracies of 89.52% and 89.40%, respectively. Nevertheless, the model all shares common limitations. These include requiring a more advanced fusion layer to effectively combine visual and spatial characteristics, increased training efficiency for the model based only on transformers, and improved information exchange between the ITP and GTP pathways.

In addition to the approaches above, facial expression recognition (FER) may be achieved using various models and techniques. For example, Zhang et al. [36] (2019) proposed a hybrid deep learning model called Deep Belief Network (DBN) for analyzing video sequences. Advanced characteristics from segmented clips are extracted using two convolutional neural networks (CNNs): spatial CNN for static face pictures and temporal CNN for optical flow images. The spatiotemporal characteristics at the segment level were combined into a deep fusion network to learn distinctive features simultaneously. The DBN segment-level features extracted from video sequences were combined and aggregated to provide a global feature representation of fixed length. A Support Vector Machine (SVM) encoded global video characteristics for Facial Expression Recognition (FER). The model underwent evaluation on the BAUM-1, RML, and MMI datasets, resulting in accuracies of 55.85%, 73.73%, and 71.43%, respectively.

The investigation of different approaches for Facial Expression Recognition (FER), such as transfer learning, face graphs using neural networks, and transformer models, showcases notable progress in capturing facial characteristics and enhancing accuracy in recognition. These novel methods tackle issues such as insufficient data, obstructions, and variations in posture, showcasing their capacity to improve Facial Expression Recognition (FER). Nevertheless, they also encounter challenges such as the intricacy of computations and the need for enhanced model integration, highlighting potential subjects for future investigation. The summary of Transfer Learning and Other Methods for Facial Expression Recognition (FER) Approaches as presented in Table 4.

Table 4 Summary of Transfer Learning and Other Methods for Facial Expression Recognition (FER) Approaches

Paper	**Dataset**	**Accuracy**	**Key Technique**	**Limitation**
Deep learning-based facial emotion recognition for human-computer interaction applications [29]	CK+	MobileNet: 98.5%, Resnet50: 97.7%, VGG19: 96%, Inception V3: 94.2%	Transfer learning with pre-trained networks (Resnet50, VGG19, Inception V3, MobileNet)	Small dataset, potential overfitting, dependency on pre-trained models
Facial Emotion Recognition Using Transfer Learning [30]	CK+, NVIE	NVIE: 99.3%, CK+: 98.3%	Hybrid approach combining transfer learning with SVM	Dependency on dataset availability and diversity
Cross-Modal Dynamic Transfer Learning for Multimodal Emotion Recognition [31]	CMU-MOSEI, IEMOCAP	IEMOCAP: 11.03% increase over TAILOR	Cross-modal dynamic transfer learning (CDaT)	Reliance on reliable confidence metrics for transfer learning
Facial Expression Recognition in the Wild Using Face Graph and	Lab-controlled, Wild dataset	89.54% (overall)	Face Graph with Attention mechanism	Reduction in accuracy in rapid facial pose or severe occlusions

Attention [32]				
Facial Expression Recognition Based on Graph Neural Network [33]	JAFFE, FER2013, CK+	95.85% (CK+)	Graph Convolutional Neural Network (GCNN) with facial landmarks	Challenges in distinguishing similar expressions, occasional misclassifications
FaceTopoNet: Face Graphs and Dynamic Topology Learning [34]	AffectNet, FER2013	AffectNet: 57.3%, FER2013: 75.8%	Dynamic topology learning with face graphs	Susceptibility to occlusions, imbalanced data, computational complexity
Appearance and Geometry Transformer (AGT) [35]	RAF-DB, FERplus	RAF-DB: 89.52%, FERplus: 89.40%	Self-attention with transformers, combining visual and spatial characteristics	Need for advanced fusion layer, improved training efficiency
Learning Affective Video Features for Facial Expression Recognition via Hybrid Deep Learning [36]	BAUM-1s, RML, MMI	BAUM-1s: 55.85%, RML: 73.73%, MMI: 71.43%	Hybrid deep learning with CNNs and Deep Belief Network	Lower accuracy for certain expressions, computational complexity

2.5 FER using Signal-based Methods

The integration of signal processing approaches via electroencephalogram (EEG) information has dramatically enhanced Facial Expression Recognition (FER). These approaches provide a robust and reliable alternative to conventional techniques that rely on images.

Using Continuous Wavelet Transform (CWT) to convert EEG signals into scalogram images, the study by Muzaffer Aslan [37] (2022) demonstrates a novel approach where a pre-trained GoogLeNet model extracts deep feature from these images. These features are then classified using k-nearest Neighbor (k-NN), Support Vector Machine (SVM), and Extreme Learning Machine (ELM). When applied to the GAMEEMO dataset, the approach demonstrated impressive accuracy rates of 98.78% using SVM, 98.53% using k-NN, and 98.41% using ELM. Although the work has been successful, it acknowledges that the computing time needed is lengthy because of the many EEG channels. It proposes more investigation into channel selection to enhance processing time.

Building on the idea of integrating multiple modalities, the paper by Erhan Ekmekcioglu, and colleagues [38] (2020) introduces a novel system that combines facial expressions, galvanic skin response (GSR), and EEG data using a hybrid fusion strategy. The feature extraction process utilizes pre-trained CNN models, namely InceptionResnetV2. Using decision tree to combine feature-level and decision-level fusion to enhance recognition accuracy. The algorithm had a peak accuracy of 91.5% when tested on the DEAP dataset.

Combining Wavelet Packet Transform (WPT) with multi-level Generative Adversarial Networks (GANs), the study by Jun Shao et al. [39] (2022) introduces an innovative approach to face aging. The technique includes an innovative multi-level generator and discriminators, each with an age and gender classifier. This removes the need for a pre-trained module that preserves identification, resulting in reduced training time and improved identity preservation. The results of extensive trials conducted on the MORPH and CACD datasets demonstrate that the proposed model surpasses current

state-of-the-art approaches in terms of both aging accuracy and identity verification. Although the model has achieved notable accomplishments, it has its limits. These include inaccuracies in labeling and difficulties in accurately synthesizing distinct characteristics of aging hair in older age groups.

Further enhancing the understanding of EEG-based emotion recognition, the review by MD. Rabiul Islam et al. [40] (2022) categorize approaches into deep learning-based and shallow learning-based systems. Deep learning techniques often yield high accuracies, frequently over 80%, and some models even attain accuracy rates as high as 99.72%. On the other hand, less advanced learning methods like Support Vector Machines (SVM), k-nearest Neighbors (kNN), and Random Forest (RF) often provide lower levels of accuracy, with an average of around 76.49%. This review emphasizes the complexities and computational demands of deep learning models and the variability in EEG data acquisition and preprocessing methods that affect the generalizability of these systems.

Signal-based approaches for facial expression recognition (FER), especially those that use electroencephalogram (EEG) data, show potential improvements in accurately recognizing emotions. By using sophisticated deep learning methods and tackling computational obstacles, these approaches may significantly enhance the resilience and versatility of FER systems in many fields, such as human-computer interaction and emotional computing. The summary of Signal-Based Facial Expression Recognition (FER) Approaches as presented in Table 5.

Table 5 Summary of Signal-Based Facial Expression Recognition (FER) Approaches

Paper	**Dataset**	**Accuracy**	**Key Technique**	**Limitation**
Emotion Recognition From EEG Signal Focusing on Deep Learning and Shallow Learning Techniques [37]	DEAP, SEED	Up to 99.72% for deep learning methods, ~76.49% for shallow learning methods	Deep learning (CNN, DBN, RNN) and shallow learning (SVM, kNN, RF)	Complexity and computational demands of deep learning models, variability in EEG data acquisition and preprocessing
CNN based efficient approach for emotion recognition [38]	GAMEEMO	98.78% (SVM), 98.53% (k-NN), 98.41% (ELM)	EEG to scalogram images using CWT, GoogleLeNet-based deep learning	Computational time due to large number of EEG channels, need for channel selection optimization
Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion [39]	LUMED-2, DEAP	81.2% (LUMED-2, one-subject-out), 91.5% (DEAP, one-subject-out)	Hybrid fusion of facial expressions, GSR, and EEG data using pretrained CNN and decision tree	Assumption of no deceptive facial expressions, challenge of integrating multiple modalities, need

				for accurate signal alignment
Wavelet-based Multi-Level Generative Adversarial Networks for Face Aging [40]	MORPH, CACD	Mean age differences range from 0.06 to 1.27 (MORPH), Identity verification confidence 95.29%-96.21% (MORPH)	Wavelet Packet Transform (WPT), multi-level GANs with age and gender classifier	Labeling errors, challenges in synthesizing prominent aging hair, need for additional facial attributes

3. GENERAL PROCESS FOR FER USING DEEP LEARNING

Facial Emotion Recognition (FER) through Deep Learning involves systematically comprehending and classifying facial expressions accurately. Each method has its own advantages and limitations; however, FER problems are generally solved through specific steps. The authors suggest the following steps outline the general workflow for FER as in Figure 3.

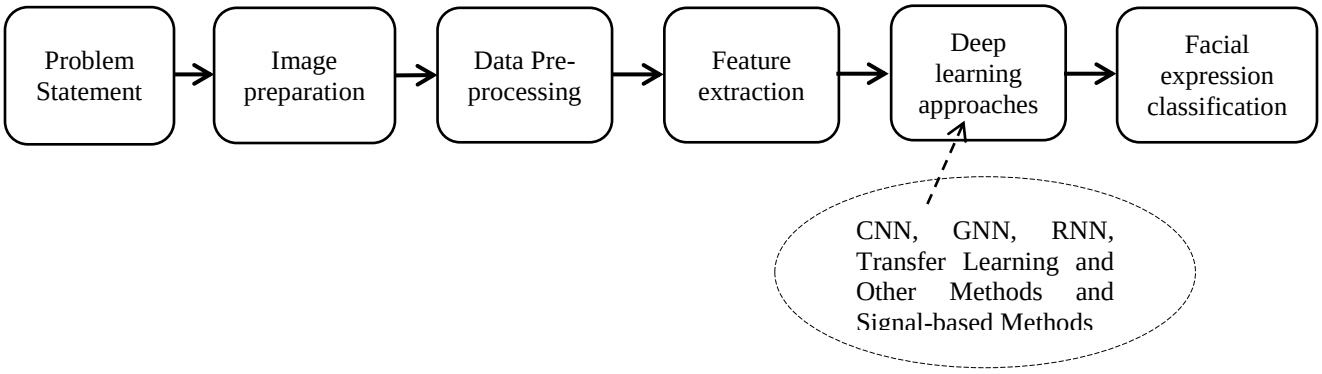


Figure 3. General process for FER using deep learning models

Step 1. Problem Statement

In the initial phase of Facial Emotion Recognition (FER), it is essential to define the problem by understanding nuances in facial expressions and potential dataset variations. Input involves diverse image types sourced from various devices, from grayscale to RGB. Output specifications, whether numerical or textual, need clarification. Challenges encompass lighting variations, facial poses, and real-time processing requirements. This precise problem definition sets the groundwork for a practical FER approach.

Step 2. Image preparation

Dataset preparation is an essential stage in FER. Choosing relevant datasets with various facial expressions is critical for developing a robust model. We can easily find data using Google Images, Kaggle, or self-prepared data. Commonly used data sets are CK+, FER2013, JAFFE, CK+, and OULU-CASIA.

Step 3. Data Pre-processing

Data pre-processing is an essential phase in facial expression classification, as we only want the part that has the necessary information. In addition, images in the datasets sometimes differ from the same type and pattern. Therefore, it is crucial to pre-process the datasets before training to make them more efficient.

Prepare the data for model training by applying various pre-processing techniques. These tasks involve resizing images, normalizing pixel values, and addressing imbalances in the dataset. Techniques like image augmentation can artificially increase the training set diversity and improve model generalization.

Step 4. Feature extraction

In CNN-based FER, the convolutional layers execute feature extraction implicitly. These layers develop hierarchical representations of facial traits, including low-level details and high-level patterns. Additional techniques like edge detection or spatial attention processes can improve feature extraction in specific contexts.

Step 5. Deep learning approaches

Choose an appropriate approach for FER based on the specific requirements of the task. Common approaches include using Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), Recurrent Neural Networks (RNNs), and Transfer Learning. The selection depends on factors like the available data, computational resources, and the desired level of model interpretability. If the problem requires comparing the performance of numerous models, we can employ different architectures to determine the most optimal model by assessing their statistics.

Step 6. Facial expression classification

The classification stage trains the model to classify facial expressions into specified emotion categories. Depending on the approach, researchers may employ various classification techniques. For instance, CNNs commonly utilize Softmax activation in the output layer for multi-class classification, while GANs may employ discriminators to classify generated images.

4. CONCLUSION

In this paper, the authors have conducted a survey and analysis of deep learning models used for the FER problem, thereby categorizing commonly used models for this problem, including Convolutional Neural Networks, Generative Adversarial Networks, Recurrent Neural Networks, Transfer Learning, Other Methods, and Signal-based Methods. For each method, the paper summarizes the principles of implementation, datasets, accuracy, key techniques, and limitations. This provides readers with an overview of the research landscape in this field and readily available resources to implement and develop new methods for the FER problem. From the related research, the paper also proposes a general process for building and developing suitable methods for the FER problem. In the following research, the authors will develop a specific deep learning model and compare it with the methods surveyed in this paper.

REFERENCES

[1] S. Singh and F. Nasoz, "Facial Expression Recognition with Convolutional Neural Networks," 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2020, pp. 0324–0328, https://doi.org/10.1109/CCWC47524.2020.9031283.

[2] R. Jaiswal, "Facial Expression Classification Using Convolutional Neural Networking and Its Applications," 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), RUPNAGAR, India, 2020, pp. 437-442, https://doi.org/10.1109/ICIIS51140.2020.9342664.

[3] Shao, Jie, and Qian Ye, "Three Convolutional Neural Network Models for Facial Expression Recognition in the Wild." Neurocomputing, vol. 355, Aug. 2019, pp. 82–92, https://doi.org/10.1016/j.neucom.2019.05.005.

[4] Jain, Deepak Kumar, et al, "Extended Deep Neural Network for Facial Emotion Recognition." Pattern Recognition Letters, vol. 120, Apr. 2019, pp. 69–74, https://doi.org/10.1016/j.patrec.2019.01.008.

[5] K. Mohan, A. Seal, O. Krejcar and A. Yazidi, "Facial Expression Recognition Using Local Gravitational Force Descriptor-Based Deep Convolution Neural Networks," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1–12, 2021, Art no. 5003512, https://doi.org/10.1109/TIM.2020.3031835.

[6] Min Wu, et al. "Two-Stage Fuzzy Fusion Based-Convolution Neural Network for Dynamic Emotion Recognition." IEEE Transactions on Affective Computing, vol. 13, no. 2, Apr. 2022, pp. 805–17, https://doi.org/10.1109/taffc.2020.2966440.

[7] Choi Jae Young, and Bumshik Lee. "Combining Deep Convolutional Neural Networks With Stochastic Ensemble Weight Optimization for Facial Expression Recognition in the Wild." IEEE Transactions on Multimedia, vol. 25, Jan. 2023, pp. 100–11, https://doi.org/10.1109/tmm.2021.3121547.

[8] Jianghai Lan, et al. "Expression Recognition Based on Multi-Regional Coordinate Attention Residuals." IEEE Access, vol. 11, Jan. 2023, pp. 63863–73, https://doi.org/10.1109/access.2023.3285781.

[9] Sun-Hee Kim, et al. "Facial Expression Recognition Using a Temporal Ensemble of Multi-Level Convolutional Neural Networks." IEEE Transactions on Affective Computing, vol. 13, no. 1, Jan. 2022, pp. 226–37, https://doi.org/10.1109/taffc.2019.2946540.

[10] Ying He, et al. "Facial Expression Recognition Using Hierarchical Features With Three-Channel Convolutional Neural Network." IEEE Access, vol. 11, Jan. 2023, pp. 84785–94, https://doi.org/10.1109/access.2023.3303402.

[11] Wu Xuemei, et al. "FER-CHC: Facial expression recognition with cross-hierarchy contrast." Applied Soft Computing, vol. 145, Sept. 2023, p. 110530, https://doi.org/10.1016/j.asoc.2023.110530.

[12] Yinggang He, et al. “Facial Expression Recognition Using Multi-Branch Attention Convolutional Neural Network.” IEEE Access, vol. 11, Jan. 2023, pp. 1244–53, https://doi.org/10.1109/access.2022.3233362.

[13] Zhifei Han, et al. “Emotion Recognition and Management in the Tourism Industry During Emergency Events Using Improved Convolutional Neural Network.” IEEE Access, Jan. 2024, p. 1, https://doi.org/10.1109/access.2024.3370431.

[14] Han, Ziyang, et al., “Face Merged Generative Adversarial Network with Tripartite Adversaries.” Neurocomputing, vol. 368, Nov. 2019, pp. 188–96, https://doi.org/10.1016/j.neucom.2019.08.049.

[15] Dharanya V., et al., “Facial Expression Recognition Through Person-wise Regeneration of Expressions Using Auxiliary Classifier Generative Adversarial Network (AC-GAN) Based Model.” Journal of Visual Communication and Image Representation, vol. 77, May 2021, p. 103110, https://doi.org/10.1016/j.jvcir.2021.103110.

[16] L. Yang, Y. Tian, Y. Song, N. Yang, K. Ma, and L. Xie, “A Novel Feature Separation Model exchange-GAN for Facial Expression Recognition,” Knowledge-Based Systems, vol. 204, p. 106217, Sep. 2020, https://doi.org/10.1016/j.knosys.2020.106217.

[17] K. Ali and C. E. Hughes, "Facial Expression Recognition By Using a Disentangled Identity-Invariant Expression Representation," 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 9460–9467, https://doi.org/10.1109/ICPR48806.2021.9412172.

[18] Yifan Xia, et al. “Local and Global Perception Generative Adversarial Network for Facial Expression Synthesis.” IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 3, Mar. 2022, pp. 1443–52. https://doi.org/10.1109/tcsvt.2021.3074032.

[19] Zhe Sun, et al. “A Discriminatively Deep Fusion Approach With Improved Conditional GAN (im-cGAN) for Facial Expression Recognition.” Pattern Recognition, vol. 135, Mar. 2023, p. 109157. https://doi.org/10.1016/j.patcog.2022.109157.

[20] Fangzheng Huang, et al. “Cyclic Style Generative Adversarial Network for Near Infrared and Visible Light Face Recognition.” Applied Soft Computing, vol. 150, Jan. 2024, p. 111096. https://doi.org/10.1016/j.asoc.2023.111096.

[21] Abhishek Gangwar, et al. “Triple-BigGAN: Semi-supervised Generative Adversarial Networks for Image Synthesis and Classification on Sexual Facial Expression Recognition.” Neurocomputing, vol. 528, Apr. 2023, pp. 200–16. https://doi.org/10.1016/j.neucom.2023.01.027.

[22] Daeha Kim, et al. “Towards the Adversarial Robustness of Facial Expression Recognition: Facial Attention-aware Adversarial Training.” Neurocomputing, vol. 584, June 2024, p. 127588. https://doi.org/10.1016/j.neucom.2024.127588.

[23] A. S. Rokkones, M. Z. Uddin and J. Torresen, "Facial Expression Recognition Using Robust Local Directional Strength Pattern Features and Recurrent Neural Network," 2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin), Berlin, Germany, 2019, pp. 283-288, https://doi.org/10.1109/ICCE-Berlin47944.2019.8966234.

[24] H. Liu, J. Zeng and S. Shan, "Facial Expression Recognition for In-the-wild Videos," 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG, 2020), Buenos Aires, Argentina, 2020, pp. 615-618, https://doi.org/10.1109/FG47880.2020.00102.

[25] Pau Rodriguez, et al. “Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification.” IEEE Transactions on Cybernetics, vol. 52, no. 5, May 2022, pp. 3314–24. https://doi.org/10.1109/tcyb.2017.2662199

[26] Manalu, Haposan Vincentius, and Achmad Pratama Rifai. “Detection of Human Emotions Through Facial Expressions Using Hybrid Convolutional Neural Network-recurrent Neural Network Algorithm.” Intelligent Systems With Applications, vol. 21, Mar. 2024, p. 200339. https://doi.org/10.1016/j.iswa.2024.200339

[27] Wissam J. Baddar, Sangmin Lee, et al. “On-the-Fly Facial Expression Prediction Using LSTM Encoded Appearance-Suppressed Dynamics.” IEEE Transactions on Affective Computing, vol. 13, no. 1, Jan. 2022, pp. 159–74. https://doi.org/10.1109/taffc.2019.2957465.

[28] Tong Zhang, et al. “Spatial–Temporal Recurrent Neural Network for Emotion Recognition.” IEEE Transactions on Cybernetics, vol. 49, no. 3, Mar. 2019, pp. 839–47. https://doi.org/10.1109/tcyb.2017.2788081.

[29] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, “Deep learning-based facial emotion recognition for human–computer interaction applications,” Neural Computing and Applications, vol. 35, no. 32, pp. 23311–23328, Apr. 2021, https://doi.org/10.1007/s00521-021-06012-8.

[30] S. Shaees, H. Naeem, M. Arslan, M. R. Naeem, S. H. Ali and H. Aldabbas, "Facial Emotion Recognition Using Transfer Learning," 2020 International Conference on Computing and Information Technology (ICCIT-1441), Tabuk, Saudi Arabia, 2020, pp. 1–5, https://doi.org/10.1109/ICCIT-144147971.2020.9213757.

[31] Soyeon Hong, et al. “Cross-Modal Dynamic Transfer Learning for Multimodal Emotion Recognition.” IEEE Access, vol. 12, Jan. 2024, pp. 14324–33. https://doi.org/10.1109/access.2024.3356185.

[32] Hyeongjin Kim, Byoung Chul Ko, et al. “Facial Expression Recognition in the Wild Using Face Graph and Attention.” IEEE Access, vol. 11, Jan. 2023, pp. 59774–87. https://doi.org/10.1109/access.2023.3286547.

[33] X. Xu, Z. Ruan and L. Yang, "Facial Expression Recognition Based on Graph Neural Network," 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC), Beijing, China, 2020, pp. 211-214, https://doi.org/10.1109/ICIVC50857.2020.9177430.

[34] Mojtaba Kolahdouzi, et al. “FaceTopoNet: Facial Expression Recognition Using Face Topology Learning.” IEEE Transactions on Artificial Intelligence, vol. 4, no. 6, Dec. 2023, pp. 1526–39. https://doi.org/10.1109/tai.2022.3207450.

[35] Ning Sun, et al. “Appearance and Geometry Transformer for Facial Expression Recognition in the Wild.” Computers & Electrical Engineering, vol. 107, Apr. 2023, p. 108583. https://doi.org/10.1016/j.compeleceng.2023.108583.

[36] S. Zhang, X. Pan, Y. Cui, X. Zhao, and L. Liu, "Learning Affective Video Features for Facial Expression Recognition via Hybrid Deep Learning," in IEEE Access, vol. 7, pp. 32297–32304, 2019, https://doi.org/10.1109/ACCESS.2019.2901521.

[37] M. Aslan, "CNN-based efficient approach for emotion recognition," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 9, pp. 7335–7346, Oct. 2022, https://doi.org/10.1016/j.jksuci.2021.08.021.

[38] Muzaffer Aslan. “CNN Based Efficient Approach for Emotion Recognition.” Journal of King Saud University. Computer and Information Sciences/Maǧalaẗ Ǧam’aẗ Al-malīk Saud: Ùlm Al-ḥasib Wa Al-ma’lumat, vol. 34, no. 9, Oct. 2022, pp. 7335–46. https://doi.org/10.1016/j.jksuci.2021.08.021.

[39] Yucel Cimtay, et al. “Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion.” IEEE Access, vol. 8, Jan. 2020, pp. 168865–78. https://doi.org/10.1109/access.2020.3023871.

[40] Jun Shao, and Tien D. Bui. “Wavelet-based Multi-level Generative Adversarial Networks for Face Aging.” Computer Vision and Image Understanding, vol. 223, Oct. 2022, p. 103524. https://doi.org/10.1016/j.cviu.2022.103524.